Published in the Proceedings of the AAAI Workshop on Text Planning and Realization, AAAI-88, St.Paul, Minnesota, August 25, 1988, pgs.91-98; available from the AAAI Office, 445 Burgess Drive, Menlo Park, CA 94025.

Modularity in Natural Language Generation: Methodological Issues

David D. McDonald[1]
August 1988

1. Questions to Focus Research

The goal of natural language generation research, if one is a cognitive scientist trained in artificial intelligence, is to develop a computational theory. Following Marr (1982), a computational theory of a cognitive process provides answers to problems posed by the fundamental nature of the task. Marr's theories of steriopsis, for example, addressed the problem of how distance to objects can be calculated given that a three-dimensional world was perceived via projection onto an only two-dimensional representation (the retina). Our comparable problem in natural language generation is to understand how people can efficiently convert information and intent that is couched in a relational and holistic medium (the mind) into a symbolic, low-bandwidth serial form (language) for comprehension within a social group.

To develop a computational theory, one needs first a framework to define the shape of what one takes a theory to be, and then one needs a set of questions---natural phenomena that require explanation---to focus the day to day research. In our case, the theoretical framework has been initially laid out in preliminary form in a paper presented at the Third International Workshop on Language Generation (McDonald, Meteer & Pustejovsky 1987). That paper focused on how the demand for efficient operation in the

[1] Author's present address: 9 Whittier St. #2, Cambridge, MA 02140 phone (617) 661-4514

face of initial situations of varying familiarity imposes architectural consequences on a generator. We concluded that the criteria for dividing generation into components must be rationalized, and that no simple decomposition is likely to be theoretically satisfying or efficient. The other part of a framework for a computational theory, the set of focusing questions, are the subject of this position paper.

The questions you ask define your research perspective: the kinds of phenomena you take to be interesting, the kinds of generalizations you seek in your theories. This perspective biases your methodology, leads you to investigate certain issues rather than others, and predisposes the architecture that you develop to take particular forms. Research is always guided in this way, even when the questions being pursued are not consciously articulated. Being explicit about one's fundamental goals sharpens the research's focus and enormously expedites efforts to compare work done by different projects, especially in issues of methodology.

1.1 Question for realization

I will begin illustrating this notion of "research-guiding questions" by looking at the sorts of questions that have guided the development of Mumble's realization component (McDonald 1984). Generally speaking, and for the purposes of this workshop, the term "realization" has been used broadly to refer to the process of carrying out the directives of the text planner to produce the actual stream of words (properly including the sounds and intonation as well). "Realization" also has other technical meanings within our field, for instance within Systemic Functional Linguistics, and it can be taken to simply mean the process of "making actual" some plan or specification. Under this more general interpretation of realization, the generation process might have any number of different "realization processes" within it.

My colleagues and I talk in terms of a "linguistic realization component" (in contrast with other components for text planning and the underlying reasoning system that motivates the whole generator in the first place). We further decompose linguistic realization into four interleaved processes operating over three representational levels. These processes and levels have been grouped together and called a component primarily because they are collectively responsible for the lion's share of the linguistic processing in generation; secondarily, we havd imagined that the processes in our linguistic realization component might well have a very different architecture and mode of operation than that of other "components" (a view that we subsequently dropped).

Against this background then, I believe that research on realization should be aimed at answering questions like these:

- Why is the process so fast ?
- Why do people never make grammatical mistakes ? (ignoring restarts and speech errors as independently accounted for) [2]
- Why should mixed-form exchange errors strand grammatical morphemes? [3]

The most satisfying answers to such questions will always be structural rather than stipulative. A stipulative answer is one where the designers wrote the rules of the system in just the right way to consciously insure that the desired behavior was achieved. An structural answer is one where the rule-writing notations provided were such that it was impossible to write rules where the behavior was violated. Structural answers follow from architectural principles embedded in the design of the virtual machine posited by one's computational theory of realization--*principles which make it impossible for the system to behave otherwise.*

The realization component Mumble-86 (Meteer et. al 1987), for example, answers the question of inescapable grammaticality by employing an architecture in which the grammatical properties of a context must be established before its content can be syntactically realized; the manifestation of these properties as active constraints insures that only grammatical realizations are ever considered.

1.2 Questions for text planning

By comparison to research on realization, which has been guided by models from conventional linguistics and from other computational treatments of grammar, text

---

2 What constitutes a "grammatical mistake" is, of course, a matter of technical definition. For example Fay (1977) views all grammatical mistakes as speech errors and uses them as empirical evidence that a classically transformational system is being used in generation, and Linde (1979) argues that the common grammatical mistake of using singular markings on a verb with a plural subject may reflect an across-the-board shift in the English language. Regardless of the particulars of the explanation, it is always better, methodologically, to make a categorical statement: if one begins with a hedge one may never look to see what could account for the apparent counter-examples.

3 This class of errors was first identified by Garrett (1975) and his students. Two content words of different parts of speech ("mixed form") exchange their position within an utterance, for example *She writes her slanting* (*... slants her writing*). Notice that the *s* that would have gone with *slants* did not move, nor did the *ing* that would have appeared with *writing*---they were "stranded".

planning research has been haphazard in its approach to problems and even to understanding what its problems are. Text planning has become the "grab bag" of generation research, accumulating all of the issues that cannot be laid off to linguistic realization or to the conceptual representation of the underlying program. In aid of this situation, I offer these two deep questions that have become the focus of my own long term projects on text planning.

What is the relationship between thought and words? [4]

Presumably the texts we utter reflect elements of our state of mind at the time we are speaking. Is the relationship direct and compositional, involving a "language of thought" (Fodor 1975) comprised of units that have a causal, constructive relationship to the surface units of our natural language? Or is it indirect and holistic, with, say, perceived situation types controlling the selection of large segments of an utterance simultaneously?

What makes utterances coherent?

Nearly all of the time what we say is germane and carries the collective situation forward along lines mutually recognized and sanctioned by speaker and hearer. When this coherence is not present, either one or another of the parties to the conversation for some reason has an incompatible idea of what the situation is, or, to give an extreme case, the speaker is schizophrenic and living in an alternative reality.

1.3 A question at the interface between text planning and realization

A third question for text planning, below, is the most immediately relevant to the topic of this workshop, since it can provide a rationale to govern the vocabulary of relations and categories that is used to communicate between text planning and linguistic realization. It is also the one that we are most likely to be able to answer soon given the state of today's research.

---

[4] While this is the simplest way to state the question, it would be better put as the relationship between thought and "linguistic resources", intending by that term to encompass the entire range of means available in natural language for expressing language besides words, e.g. grammatical morphemes, syntactic constructions, stylistic options, intonation, large-scale rhetorical organizations, idioms and fixed phrases, etc. "Language as resources" is a natural perspective to take when viewing it as constructions assembled under the direction of a planner. One also sees this view of language in the work of linguists such as Len Talmy (1987).

Why are messages realizable?

Only rarely do people "talk themselves into a corner" (if it were commonplace we wouldn't remark on it when it happens). We can take this fact as evidence that the messages people plan work reliably, defining a "message" to be the specification of the information that the speaker wants communicated in a utterance. If we assume that most messages are composed dynamically to meet the needs of new situations, then we need an explanation of why it is that this information content corresponds to something that can actually be grammatically expressed given the linguistic resources available to the speaker in context (footnote four).

We can paraphrase this question as a definition of text planning:

> "Text Planning is the preparation of a specification from which Realization can effectively produce an utterance."

A definition like this stands the usual conception of text planning on its head. It ignores the customary and serious issues of how a text planner selects the information to be communicated, orchestrates it to give it an effective form, and chooses the linguistic resources that are to convey it (as shall this paper[5]). It points out instead the simple fact that *whatever a planner might want to do, it is only going to be able to do what the surface grammar and lexicon of the language let it.*

This is a very serious challenge to a conventional, two component generation architecture. How is a text planner to know what it will be allowed to do without it incorporating essentially all of the knowledge of the realization component;[6] and if the knowledge is shared, on what basis are we to say that there are two components instead of one? The proper response to this dilemma, it seems to me, is to be suspicious of the two component architecture that we have fallen into---it may be creating more problems than it solves. We should seriously consider alternative decompositions of the generation process, which in turn means that we should look carefully at our criteria for dividing a process into modules.

It is not unreasonable to expect that the "planning" done in generation may involve extremely specialized methods and that its links to "realization" may be equally specialized. The two aspects of the generator grew up together (as it were) during the

---

5 Initial work on these questions by the author and his colleagues is discussed in McDonald & Meteer unpublished and in McDonald 1988.

6 For an insightful exploration of this problem and a proposed solution, see Meteer, this volume.

course of human language's evolution. We should expect them to have coordinated, potentially unique designs, rather than to automatically fall into any of the traditional artificial intelligence molds for planning/execution such as have been applied to conscious intellectual tasks such as playing chess or organizing a day's shopping.

2. Criteria for partitioning the generation process

Now that we have invited the lines between planning and realization to be perhaps arbitrarily blurred, we must establish some independent criteria by which a theory should define independent processing modules and intermediate representations. For this, we should go back to our fundamental view of generation as a process of decision-making. Abstractly, we can think of decisions as falling into specifiable classes that are established by the particular theory one has adopted. These classes may be defined in terms of four criterial design dimensions:

(1) **What is decided** We would all agree that it is a different matter to chose a content word than to choose what tense to use. In a case like this where the things being chosen are of different kinds, the decisions could in principle be made at different times and with different preconditions and consequences. (Of course there is no requirement that everything of a given kind be handled by the same component. Should one elect not to do this, however, then there must be some additional factors, such as one of the next three criteria, that provides a consistent rationale for using different components.) For other surface resources less obviously different than words and tenses, it may only be possible to definitively parcel them out after one has established what linguistic theory is to be used in the generator, using the theory to justify the division of resources into classes; but any theoretical distinction could, in principle, be used to motivate a division of the generator into components. For example our collective impression that deciding on content is distinct from deciding on surface resources is mirrored in our initial assumptions about the divisions between text planning and linguistic realization.

(2) **The reference knowledge** Decisions are made on the basis of knowledge about what alternatives are possible, how an alternative's applicability depends upon context, and on what the consequences of a given choice may be. This long term "reference knowledge" is drawn on as choices are deliberated in order to provide the needed decision criteria. Reference knowledge bears on component design once we appreciate that all of a generator's resource knowledge need not be equally accessible

at all times. Knowledge about grammatical constraints on surface order, for example, can be ignored when making choices about how to express the manner of a motion. Differences like this are a natural basis for dividing generation into processing stages: Whenever some body of knowledge is not needed it can be omitted from the affected processing component along with any structural substrate needed to support it. Differences in the linguistic theory being used, because they can imply differences in the organization and granularity of a given element of the knowledge, also bear on this kind of modularity.

(3) **The representation of the results** Most of the time when a generator makes a decision it cannot immediately act on it. A decision may fix several aspects of the (eventual) surface utterance simultaneously, for example, and even if the first (leftmost) of them is precisely at the present point of speech, the others will have to be somehow remembered until their turn in the utterance is reached. This "memory" must be sustained via a structure couched in some kind of representation (note that changes to the state of some specialized processor would do as well). Processing theories can use these structures to good advantage: In McDonald, Meteer, & Pustejovsky 1987 we argue that the most efficient architectures will accumulate the results of one component's decisions in the structure that controls the next component's operation; the most common alternative is the feature systems employed in systemic and unification grammars. Contributing to a common output representation is a clear rationale for selecting the decision classes to go into the same component.

(4) **The representation of the information being manipulated** Equally clear as a rationale for dividing generation into components is the form of the substrate that a class of decisions is working from: surface structure, the state of the underlying program, discourse-level schemas, etc. Finer breakdowns of what one is manipulating are possible too: Intra-clausal phrase order is decided by looking at the properties of the verb and arguments being ordered, but not necessarily at the surface structure to either side; explicit vs. inferential communication of a unit of information requires awareness of what information can be carried by the available words, but may need only a very shallow, non-linguistic representation of the information that has been communicated thus far. Components can often be defined simply by the structures they operate over.

To summarize, we can see that once we accept the idea that planning and realization may be highly specialized and coordinated in their designs, the possibilities for how they may be divided into components and interleaved become quite large and may be adjudicated along several complementary dimensions.

As an engineering matter, one can always pull off a design where no processing lines are ever drawn, jumping back and forth along the dimensions (perhaps by island-driving) as circumstances warrant. This, however, would constitute a null theory of processing architecture for generation. Methodologically it is to be taken only as a last resort after all stronger theories have been proven impossible or ungainly. Assuming, as I do, that one's goal is a computational theory of generation that can in principle be applied to people (work in "theoretical psycholinguistics" as it were), then it is always better to hypothesize a division into distinct stages and processing modules, following lines like those sketched above. One then looks to see what difficulties the hypothesis may engender and varies or weakens the division boundaries as needs be. Proving a strong hypothesis wrong always leads to greater insights for the next attempt; choosing the null hypothesis from the start gains nothing.

3.2 Modularity and processing algorithms

We should now return to the specific issue of this paper: how does the text planner know what the realization component will let it do---what does it need to know (or implicitly comply with) in order to assemble a message that can be realized? We have just seen that the first thing to consider in approaching this question is whether we have drawn the lines between our components at well motivated places. We have questioned whether it is reasonable to think in terms of one seamless text planning component producing its message all at once (say for an entire sentence or paragraph) and passing it as a whole to one seamless realization component.

Leaving a definitive answer to this question to later papers, we should take up another part of the problem of assuring a message's expressibility (to use Meteer's term), which is the question of what processing algorithm is to be used. One could insure expressibility by using an algorithm based on lookahead: alternative messages would be assembled and their realizations monitored, backing up whenever a problem was encountered. However it is hard to see how a "generate and test" design could ever be made efficient, and if we know anything about the processing characteristics of human speech we know that it is fast and shows no apparent influences attributable to utterance length or syntactic complexity. Consequently a generate and test design is an unlikely

possibility for online speech when working from a cognitive perspective. (Polished writing is another matter entirely.)

An alternative style of algorithm would be a least commitment strategy. Under this kind of processing algorithm, constraint propagation and progressive refinement are used to ensure that (1) individual decisions are not made before all the information relevant to their choice has been determined, and (2) that the choices decided upon are of a character and grain size such that they are not establishing more of the utterance than the evidence at that point actually warrants, making it easier to interleave the decisions and keep the processing indelible (i.e. all decisions add to each other monotonically, without backup or throwing out established choices). This is the customary style today among researchers interested in algorithmic efficiency, and has taken many different forms (for example Hovy 1987, Jacobs 1987, McDonald 1984, Mellish 1988).

A processing algorithm based on least commitment with constraint propagation can certainly be made to be indelible. However, it is still not the most efficient processing algorithm that we can imagine. Given that the human language faculty is a highly evolved, integrated system, we should seriously consider using highly specialized algorithms in our models, which might be completely inapplicable to tasks other than generation. The problems with algorithms based on constraint propagation techniques is that they carry with them a high symbolic overhead. Descriptions of the generator's decisions and intermediate states of affairs must be created to manifest the constraints, and their examination and updating consumes resources without immediately contributing to the utterance.[7]

An efficient alternative to constraint propagation would be simply to "do the right thing"---to employ a straight-line, monotonically increasing, indelible construction process that succeeds simply by virtue of making the needed decisions in a compatible order, not passing around symbolic descriptions of constraints or making "meta-decisions" about decision orderings according to the specific situation. This is certainly an ambitious goal, but if it could be done it would arguably lead to the strongest theory.

---

[7] Descriptions may be sets of features such as those that mediate between the state of the system networks and "realization" in a typical systemic grammar. They may be property labels, specialized category systems, etc.. In all cases they are only characterizations of linguistic and conceptual elements, not real elements themselves (depending of course on one's linguistic theory and presumptions about the psychological status of its abstractions).

As a step towards a "just do the right thing" algorithm, my colleagues and I are exploring designs where the generation process is broken down into a great many strictly ordered, pipelined, and contextually interleaved processing stages (order of a dozen). Each of these stages is responsible for a relatively small amount of the total processing. For example there are likely to be individual processes for such things as selecting the theta grid (argument pattern) of a phrase's lexical head, applying perspective to select a word, selecting among alternative factorings of a complex idea according to which components have been made salient by recent mental actions, dictating the most abstract representation of the utterance by its head/argument/adjunct structure (see Meteer this volume), and specifying the elementary trees and adjunction patterns for the utterance's syntactic structure.

Decomposing generation into many processes, each contributing a small amount to the utterance, appears to us to be the only realistic means of achieving an indelible, efficient process with low symbolic overhead. In doing this, however, we end up dramatically blurring the text planning - realization distinction (particularly when the distinction is taken to be between deciding what to say and how to say it). Indeed, we have found that the organizing vocabulary that individual process employs (overlapping with the processes it abuts) is always a delicate blend of situational (semantic) and linguistic types, grading smoothly from the almost purely conceptual to the purely linguistic as the processes get closer to the actual utterance---an observation that is not compatible with any simplistic division of generation into just two components.

Our conclusion is that the pre-theoretical notion that there are two distinct components in a generator---one responsible for utterance content and organization and characterized as a planner, the other responsible for linguistic form and seen as carrying out or "realizing" the planner's plans---cannot be sustained any longer. There are too many benefits from moving to designs with many more than two distinctly organized components, and any move to lump the new components into two groups just to make it possible to use the old labels misses the point.

4. References

Fay, David (1977) "Transformational Errors" 12th International Congress of Linguists, Working Group on 'Slips of the Tongue and Ear', August 31 - September 2, 1977, Vienna Austria.

Fodor, Jerry (1975) **Modularity**, MIT Press.

Garrett, Merrill (1975) "The analysis of sentence production" in Bower (ed.) **The Psychology of Learning and Motivation** vol.9, Academic Press, 133-177.

Hovy, Eduard (1988) **Generating natural language under pragmatic constraints**, Lawrence Erlbaum Associates, Hillsdale, NJ.

Jacobs, Paul (1987) "Knowledge Intensive Natural Language Generation", **Artificial Intelligence**, **33**(3), 325-378.

Marr, David (1982) **Vision**, W.H.Freeman, San Francisco.

McDonald, David (1984) "Description Directed Control: its implications for natural language generation", **Computers and Mathematics 9**(1), 1983, 111-130; reprinted in Grosz, Sparck Jones, & Webber (eds.) **Readings in Natural Language Processing**, Morgan Kaufman, 1986, 519-538.

McDonald, David (1988) "On the place of words in the generation process" Proc. 4th International Workshop on Natural Language Generation, Catalina Island (Information Sciences Institute, Los Angeles), July 18-21, 1988.

McDonald, David, Marie Meteer, & James Pustejovsky (1987) "Factors contributing to efficiency in natural language generation", in Kempen (ed.) **Natural Language Generation**, Martinus Nijhoff Publishers, Boston, 1987, 159-182.

Mellish, Chris (1988) "Natural language generation from plans", in Zock & Sabah (eds.) **Advances in Natural Language Generation**, Pinter Publishers, London, 131-145.

Meteer, Marie (1988) "Defining a vocabulary for text planning", this volume.

Meteer, Marie, David McDonald, Scott Anderson, David Forster, Linda Gay, Alison Huettner, & Penelope Sibun (1987) **Mumble-86: Design and Implementation**, Technical Report 87-87, Department of Computer & Information Science, University of Massachusetts, Amherst, MA.

Talmy, Leonard (1987) "The Relation of Grammar to Cognition", in Rudzka-Ostyn (ed.) **Topics in Cognitive Linguistics**, John Benjamins.